DOCUMENT RESUME

ED 061 973 LI 003 643

AUTHOR	Hagerty, Katherine
TITLE	Abstracts as a Basis for Relevance Judgment.
INSTITUTION	Chicago Univ., Ill. Graduate Library School.
SPONS AGENCY	National Science Foundation, Washington, D.C.
REPORT NO	WP-380-5
PUB DATE	Feb 67
NOTE	36p.;(39 References); Studies in Indexing Depth and Retrieval Effectiveness
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Abstracts; Experimental Programs; *Indexing; *Information Processing; Information Retrieval; *Information Seeking; Information Systems; *Relevance (Information Retrieval)

ABSTRACT

The purpose of the study was to determine how many questions judged relevant to an article are also judged relevant using different length representations of the article, and how this proportion changes as the length of the representation changes. Several persons made relevance judgments for each of a set of articles to each of a set of questions. For each article several different length representations, titles and abstracts, were used as the basis for relevance judgments. The study showed that the title alone was quite effective in making relevance judgments and that even with a 300 word abstract, this effectiveness in terms of recall and precision, was not doubled. However, the use of an abstract did increase the recall and precision, but the results of the study do not suggest an optimal abstract length. (Author/NH)

ED 061973

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

THE

UNIVERSITY OF CHICAGO

GRADUATE LIBRARY SCHOOL

STUDIES IN INDEXING DEPTH AND RETRIEVAL EFFECTIVENESS

(NSF GN 380)

Working Paper No. 380-5

ABSTRACTS AS A BASIS FOR RELEVANCE JUDGMENT
(M.A. Thesis)

Katherine Hagerty

Feb. 1967

LI 000 643

TABLE OF CONTENTS

	Page
LIST OF TABLES	iii
LIST OF ILLUSTRATIONS	iv
INTRODUCTION	1
PREVIOUS RELATED WORK	2
GENERAL PROCEDURE	5
EXPERIMENTAL DESIGN AND DESCRIPTION OF PROCEDURE	6
RESULTS	11
DISCUSSION OF RESULTS	20
CONCLUSIONS AND SUGGESTIONS	21
APPENDIX	23
BIBLIOGRAPHY	29

LIST OF TABLES

Table		Page
1.	Average Recall for Each Person over All Articles	23
2.	Part 1: Recall--Precision	24
3.	Part 2: Recall--Precision	25
4.	Part 1: Total Correct Judgments	26
5.	Inconsistency in Judging Relevance of Questions Using Whole Article as Basis	27

LIST OF ILLUSTRATIONS

Figure — Page

1. Average Recall and Precision over All Articles: (a) for Each Person; (b) over All Persons 12

2. All Points of Recall and Precision for Each Length Representation 14

3. Improvement in Recall and Precision for Each Person and Each Length Abstract 16

4. Average Improvement in Recall and Precision for Each Length Abstract over All Articles: (a) for Each Person; (b) for All Persons 18

5. Cumulative Improvement in Recall and Precision for Each Person for Title--60 Words and Title-300 Words 28

INTRODUCTION

The study being reported was an experimental investigation of individual relevance judgments using different length representations of an article. The relevance judgment was defined as a judgment or an opinion about the probable relevance of an article to a question. Several persons made relevance judgments for each of a set of articles to each of a set of questions. For each article several different length representations, titles and abstracts, were used as the basis of the relevance judgments. Each person decided which of the questions of a question list might be relevant to an article on the basis of a representation of the article. After making relevance judgments to a question list using the representations of each article, he used the whole article as the basis for relevance judgments. An article would be judged relevant to a question if the person believed it to be relevant to the question, i.e., if he thought it had a bearing on the question or contained information to answer the question.

The purpose of the study was to attempt to find how many questions judged relevant to an article are also judged relevant using different length representations of the article, and also how this proportion changes as the length of representation changes.

PREVIOUS RELATED WORK

If a user of an information system is given a list of articles in response to his request, he may not necessarily wish to read all the articles. That is, the specific piece of information he wants may be contained in any or all of the articles listed. If he is interested only in that information and has a limited amount of time to spend for this request, he may wish to read only the most pertinent article or articles. In helping the user make relevance judgments, the article itself, is, of course, the most reliable key to a judgment of relevance. It is not, however, the quickest or simplest key to use. A brief representation of the article could be read by the user in less time than the whole article. What length representation would be most helpful to the user in general in determining the relevance of articles for his request is not known.

There are other problems associated with relevance judgments and with the interpretation of relevance. As Rees[1] points out, the user's judgments of relevance are a subjective reaction, are not consistent over time, and are not comparable to another user's judgments because documents "inform" various people differently. Taylor[2] hypothesizes that because of the difference in interpretation of the relevance of a document to a given request, each searcher will choose a different subset of documents as relevant to that question.

Goffman[3] concludes that "if relevance is defined as a measure of information conveyed by a document relative to a query," this relationship can only be determined with respect to a set of documents, not just between a document and the question. The user, in other words, will rank articles in terms of their relevance to his request; one article will be more relevant, conveying more information relative to his query, than another. This relationship is not completely stable because it is dependent on human judgments of relevance. The factors that influence these judgments

---

[1]Alan M. Rees and Tefko Saracevic, The Measurability of Relevance, Center for Documentation and Communication Research, Western Reserve University (Cleveland, Ohio: Western Reserve University, October, 1966), pp. 8-9.

[2]Robert Taylor, "The Process of Asking Questions," American Documentation, XVI (July, 1965), 393.

[3]Rees and Saracevic, p. 1.

FILMED FROM BEST AVAILABLE COPY

"are numerous and most of them are unknown to the person himself."[1] Since the influence of these factors would most likely vary on different occasions, judgments of relevance for the same user over time will probably result in inconsistency. Taube[2] accepts the "supposition that the searcher uses the term relevance with fair consistency in referring to his own responses." When Resnick and Savage[3] compared different kinds of representations of articles, their results showed a consistency of judgments for the same user over time except in the case of abstracts. Although such representations as titles or abstracts are used to make relevance judgments, it is questionable how well they can substitute for the entire document.

If a user were given abstracts one-half the length of the articles in order to judge the probable relevance of an article to his request, it might be expected that all the articles selected would be relevant because he had a quite significant amount of information about the article as the basis of his judgment. DeLucia[4] did find that students using abstracts, which were an average reduction of 49 per cent of the articles, performed as well or better than those students using the full articles in tests of comprehension.

For the simplest representation of articles titles can be used as the only basis of relevance judgments. When Resnick[5] compared the effectiveness of titles and abstracts as a basis of relevance judgments, he reported two results. First, titles and abstracts of articles were about equal in effectiveness when relevance judgments were made by a user for articles in terms of his work interest. Second, abstracts performed significantly better than titles when obtaining specific information on specific questions by relevance judgments because the abstracts contained more answers to the questions.

---

[1] Ibid., p. 10.

[2] Mortimer Taube, "A Note on the Pseudo-Mathematics of Relevance," American Documentation, XVI (April, 1965), 71.

[3] A. Resnick and T. R. Savage, "The Consistency of Human Judgments of Relevance," American Documentation, XV (April, 1964), 93-95.

[4] Al DeLucia, "Index-Abstract Evaluation and Design," American Documentation, XV (April, 1964), 121-25.

[5] A. Resnick, "Relative Effectiveness of Document Titles and Abstracts for Determining Relevance of Documents," Science, CXXXIV (October 6, 1961), 1004-1006.

In the study reported by Rath, Resnick and Savage[1] on the use of different types of representations of articles to select documents relevant to lists of questions, there were no major differences between the groups using complete text and abstracts. Also, in the experiment, the use of titles resulted in a high proportion of correct responses.

---

[1]G. J. Rath, A. Resnick, and T. R. Savage, "Comparison of Four Types of Lexical Indicators of Content," American Documentation, XII (April, 1961), 126-30.

GENERAL PROCEDURE

In the study reported here, many questions were devised from each article. In a retrieval experiment there is usually a large document collection and a collection of questions or requests. Several documents may be relevant to a request. These requests may be a collection of actual requests from a library, etc., or devised from the articles used in the experiment. This study consisted of a small document collection and a large collection of questions which were directly related to, i.e., devised from, the articles. All the articles and questions were in general concerned with the same subject. However, since many of the questions were quite specific, information pertinent to a particular question may not have been contained in every article in the collection.

In the study certain assumptions were made about relevance. (1) A relationship of probable relevance can be determined on the basis of a question and one article; articles will not be ranked. (2) Each person may relate a different subset of questions to an article because of the subjective nature of relevance. Thus, a composite list of questions relevant to an article containing only the questions agreed upon as relevant by all persons will not be compiled. (3) Since the relevance judgments may be inconsistent for the same user over time, relevance judgments for articles will be repeated to determine the degree of consistency.

FILMED FROM BEST AVAILABLE COPY

EXPERIMENTAL DESIGN AND DESCRIPTION OF PROCEDURE

The document collection for the study consisted of thirteen articles, selected from several periodicals and one book of collected short papers, dealing with the common subject of scientific communication among scientists. Of approximately equal length, the articles chosen contained as few graphs and tables as possible so the abstracting procedure would deal mainly with verbal text reduction.

The study was divided into three consecutive parts; a pilot study, Part 1 and Part 2. Much of the data was analyzed by comparison of Part 1 to Part 2 because the procedures were the same. The pilot study's data were used primarily to determine what length abstracts should be used in the next two parts. Since the number of abstracts and questions were different for the pilot study, its data were not always comparable to that of the other parts. Four articles were assigned to each of the three parts; the remaining article in the collection was used in formulating questions but was not abstracted.

The representations for these articles were titles and abstracts written for the study. The abstractors and judgers were four graduate students in library science. Since they knew the contents of the article they had abstracted for each part, they did not make relevance judgments using their own abstracts.

For each part of the study the abstractors, who had neither training nor experience in abstracting, were each assigned a different article. There were no explicit instructions on the type of abstract, informative or indicative, to be written, but it was presumed that the main concepts or coverage of the whole article would at least be indicated. How much an abstract can be informative or indicative is to a great extent dependent on both the person and the article. It is difficult to say whether results would have been different if the abstracts had conformed to a common standard. However, to determine if any one article or abstractor appeared to affect the results collected, the data were analyzed separately for each article and for each abstractor.

The abstractors followed instructions for length and wrote the successive abstracts for each part in order of their increasing length. The shorter representations were not incorporated in the successive longer ones, except for inclusion of the title in Parts 1 and 2. Each length abstract was written in reference to the whole article. It was felt that a longer abstract might provide for greater detailing of a concept or an approach than possible in the previous length abstract and that

the abstractor should have freedom of choice in wording and vocabulary in writing each length abstract. If he had been required to incorporate each length abstract in the longer one, he may have felt limited by his first choices of vocabulary and wording. It may have been preferable that each abstract be incorporated in the next longer one so that the same information was basic to each length representation of an article.

For the pilot study there were two lengths of abstract, 30 and 300 words. Generally, a 30 word abstract contains approximately 10 substantive words and this was assumed to be the minimum meaningful description. The 10-to-1 ratio was arbitrarily chosen; since most of the articles were about five pages long, a 300 word abstract was presumed to be a fairly substantial representation of the article.

But this 10-to-1 ratio does not represent the amount of increase of information in the longer abstracts. Hirayama[1] analyzed five different length abstracts of an article and concluded that the percentage of information in an abstract tends to decrease as the length of abstract increases. It is not known, however, when the amount of information necessary for certain purposes has been reached.

From the results of the pilot study, it was decided to add to the number of representations of an article to be judged. The title and another length of abstract were chosen; for Part 1 and Part 2 there were four representations of the article, title and 30, 60, and 300 word abstracts. As mentioned previously, each length abstract then also included the title of the article.

The questions used in making the relevance judgments were formulated by four persons; two faculty members and two graduate students in library science who did not participate in the abstracting and judging. Each person devised 10 questions from each of four articles, a total of 160. For three articles questions were devised by two persons. Although these questions were formulated directly from articles, those making up questions attempted to devise ones they might really ask of an information system. Moreover, they tried to use as little of the vocabulary of the article as possible so the connection between a representation and a question would not be merely a process of matching the words contained in them. The figure of 10 questions per article was somewhat arbitrarily chosen. When several questions are devised from an article, the tendency, if present, to base questions solely on the title or the main concept is minimized.

---

[1]Kenzp Hirayama, "Length of Abstract and Amount of Information," Journal of Chemical Documentation, IV (January, 1964), 9-11.

A question list was compiled for each part of the study. For the pilot study the list consisted of 40 questions directly related to the four articles used, 10 questions to each. Since the persons judging relevance knew the composition of the list, this may have influenced their judgments; they could eliminate the questions seemingly related to the article they had abstracted and try to match the others. Each of the question lists for the next two parts of the study corrected this weakness and consisted of a selection of 100 questions drawn from the entire question collection. The number of questions in a question list directly related to any of the articles abstracted for that part of the study was not the same for all four articles. This number was not made known to the judgers and the list also included questions not formulated from any one of the articles being used in that part of the experiment.

Then, relevance judgments were made using the question list and the abstracts for each part. For each question each person decided whether the article represented by, for example, the title, "Information Retrieval: State of the Art," might be relevant. An article might be relevant if an inference of probable relevance could be made from the representation of the article; proof of relevance did not have to exist in the representation. Three choices were available in the pilot study to indicate the decision of relevance, relevant, irrelevant and "unable to decide." The latter choice was seldom made, and led to more complexity in the analysis of results than was justified. Accordingly only the alternatives relevant and irrelevant were retained for the decisions of relevance in Parts 1 and 2.

Relevance judgments were made by the judgers using the text reductions in order of increasing length. Each person then determined for each article which questions were relevant using the article itself as the basis. Thus, a list of relevant and irrelevant questions, called the base list of questions, was created for each person for each article. The base list of questions was then used as the standard against which the lists of relevance judgments made with the text reductions of an article by one person were measured. After a period of two weeks, the same persons made repeat relevance judgments to provide some measure of consistency using the whole article as a basis.

Because of the nature of the results presented, the data were arranged for individuals and for articles. The individual's base list of questions for an article was used as a basis for evaluating, or measuring, that individual's relevance judgments made using a representation of that article; the term case refers to the results of this measure. That is, the results for one person for one article and one of its representations is referred to as one case. Although the majority of the data was

expressed in terms of the individual person for each case, some averages over cases such as for consistency were calculated.

Two of the measures used for evaluation were called recall and precision. These measures are often used in evaluating results of retrieval experiments on the performance of information systems. Documents are retrieved from a collection by some means being tested in response to a request. Various methods establish how many documents there are in the collection which are relevant to the request. In order to evaluate the retrieval system in terms of the user's request, it is important to see how well the system performed in retrieving relevant documents out of the collection. Recall is then the ratio of the relevant documents that were retrieved to the total number of relevant documents in the collection. If there were 10 documents relevant to a request in the collection and 5 of these were retrieved by some means, the recall would be 50 per cent. The term precision is used for the ratio of the number of relevant documents that are retrieved to the total number of documents that are retrieved, which includes the number of relevant and the number of irrelevant documents. Thus, if there were 15 documents retrieved in response to a request and 5 of these were relevant, then the precision would be 33 per cent.

If precision in a particular system was not very high and the user were aware of this, he could sort out the relevant documents from those retrieved in response to his request maybe by titles and abstracts. This would be a human judgment of relevance and it would be important to know how well this task could be performed.

In the study being reported, the purpose in calculating the results is analogous to that in a retrieval experiment. In order to see if the user can sort out the relevant documents in a list by means of a representation of the document and if he also includes some of the irrelevant material retrieved, the same terms were used. Here recall is the ratio of the number of the questions correctly judged relevant using a representation of an article to the total number of relevant questions (i.e., the list of questions created using the whole article as the basis of judgments). Precision is the ratio of the number of questions correctly judged to be relevant using a representation to the total number of questions judged relevant using that representation. The diagrams below illustrate the two uses of the terms recall and precision.

documents			questions		
relevant documents not retrieved	relevant documents that were retrieved	retrieved documents not relevant	questions really relevant not judged relevant	relevant questions judged relevant	questions judged relevant which are irrelevant
a	b	c	a	b	c

Recall in both cases is $\frac{a}{a+b}$ and precision is $\frac{b}{b+c}$.

Another measure used for evaluation was called total "correct"judgments. This means the number of judgments made using a representation of an article which matched the judgments made using the whole article. Total correct judgments includes both the number of correct judgments of relevance and the number of correct judgments of irrelevance and is expressed as a percentage.

RESULTS

For successively longer abstracts, because the amount of information available to the judger increases, one would expect recall also to increase. Most of the data for each individual does support this expectation and the average for 21 cases did show an increase. The average for the judgments made using the titles was 43 per cent recall and 51 per cent precision while the average for judgments made using the 300 word abstract was 64 per cent recall and 72 per cent precision. See figure 1. The standard deviation for the title judgments was 19.1 per cent for recall and 16.8 per cent for precision; for the 300 word abstracts the standard deviation was 14.9 per cent for recall and 15.8 per cent for precision.

For the 21 cases[1] the average number of questions judged relevant in the base list of questions was 24 out of 100. Thus, 24 per cent of the questions were considered, on the average and with a standard deviation of 6.2 per cent, to be relevant. If a random selection of the questions were made, one might expect that 24 per cent of those questions selected would be relevant ones. Thus, one may assume that if judgments of relevance for the 100 questions were made at random, then the precision expected would be 24 per cent. Because these are human judgments of relevance the possibility exists of doing worse than random. In two cases, one for a title and one for a 30 word abstract, precision was worse than random and in one case for a 30 word abstract it was no better than random. Precision of judgments for abstracts longer than 30 words was consistently better than random.

---

[1]Part 1 consists of 12 cases and Part 2 (incomplete) consists of 9 cases.

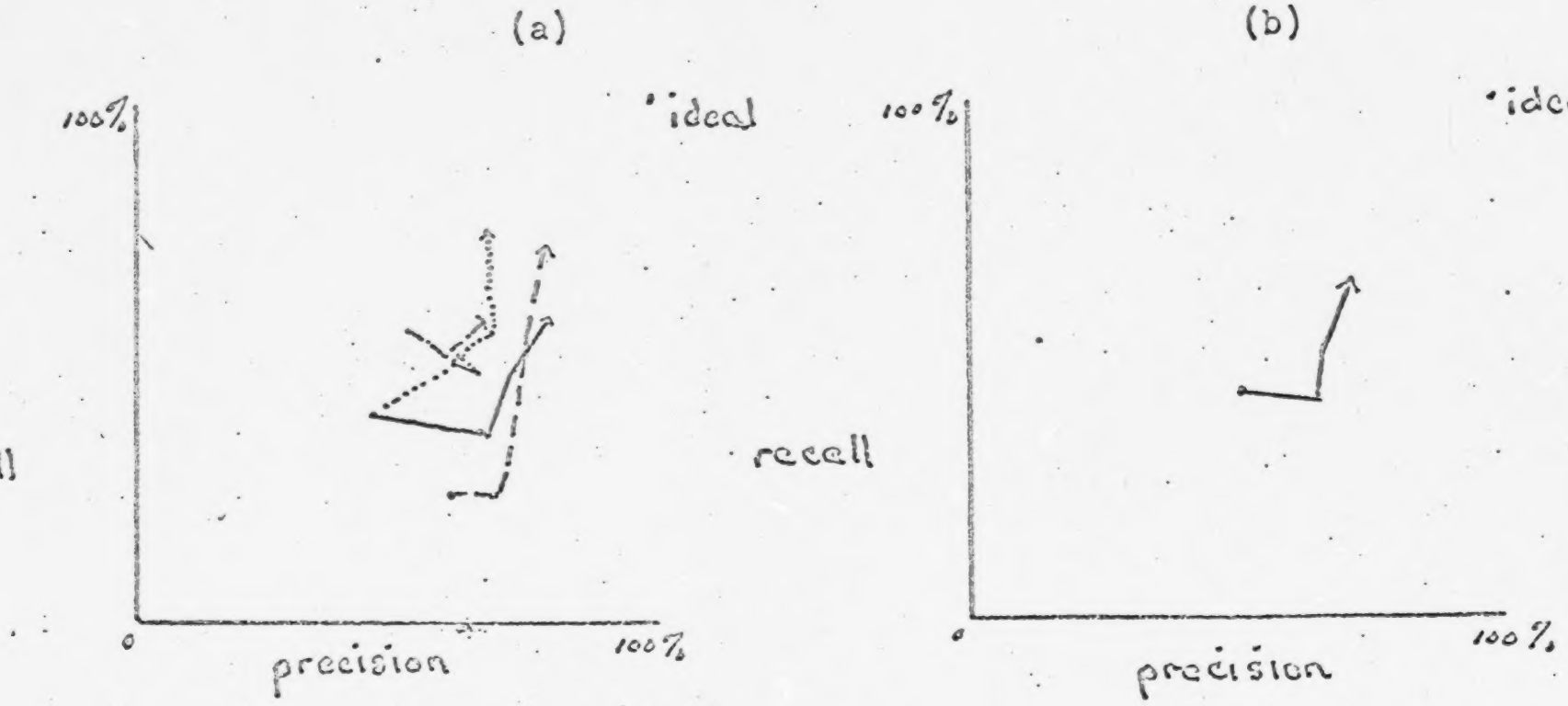
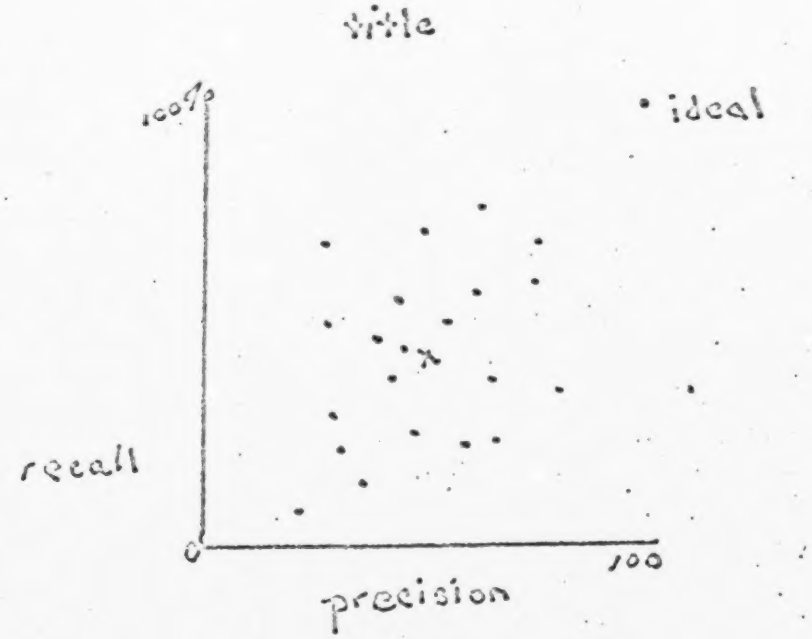
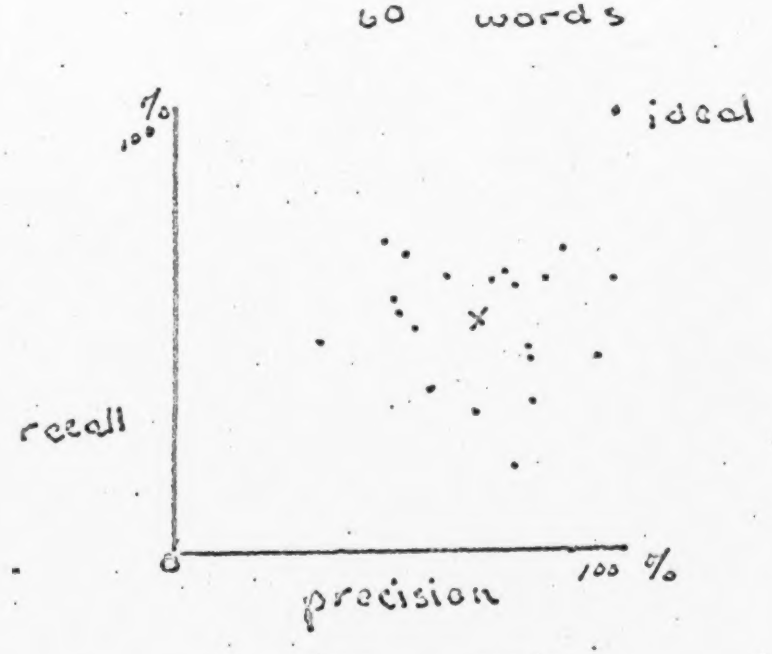
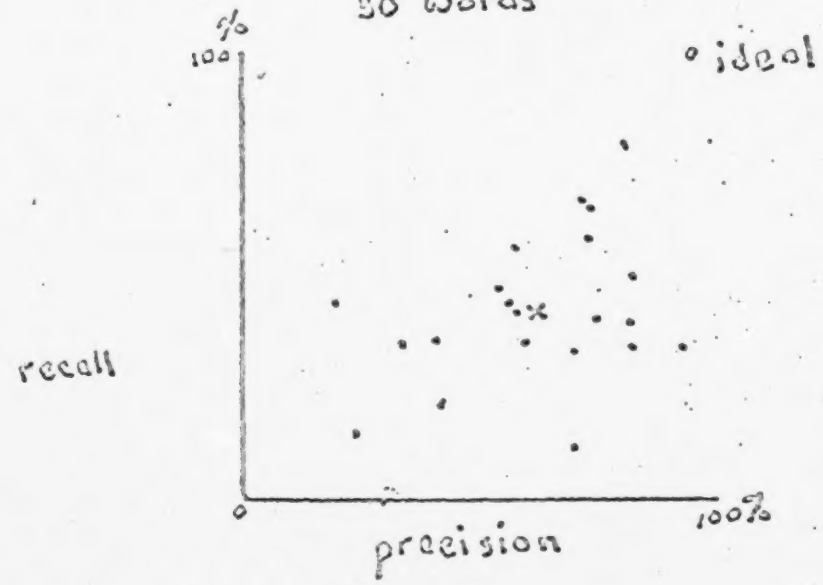


Fig. 1.--Average Recall and Precision over All Articles:
(a) For Each Person
(b) Over All Persons

There was a great dispersion of points of recall and precision and in many cases it is surprising to see how effective the title appears to be for the making of correct relevant judgments, as shown in figure 2. Although results for the title were good, the results for the 300 word abstract were apparently better in most cases, only one of 21 cases having less than 50 per cent recall and three cases having less than 50 per cent precision. For the title two cases had 50 per cent recall with thirteen cases of less than 50 per cent recall and one case had 50 per cent precision with 10 cases of less than 50 per cent precision. The average precision for the title, as mentioned (page 11), was 51 per cent precision--a little more than double the random precision of 24 per cent. But for the 300 word abstract, precision (72 per cent) was three times the random precision, a significant improvement.







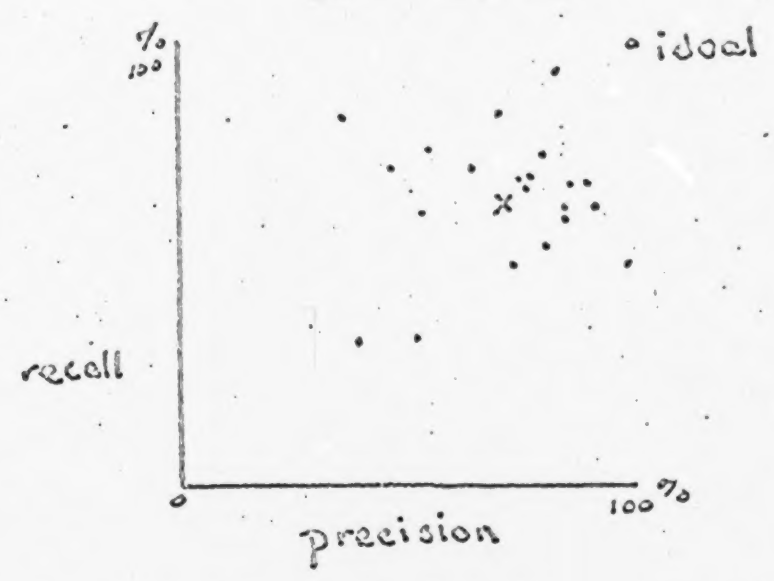
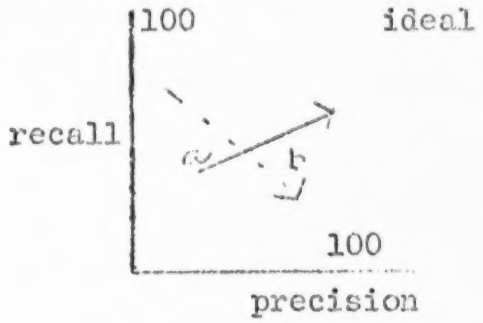


Fig. 2.--All Points of Recall and Precision for Each Length Representation.

The ideal point is 100 per cent recall and 100 per cent precision; in this situation all relevant questions are selected by the judger and all the questions selected as relevant are in fact relevant. As the length of representation is increased, one might expect that recall and precision would approach the ideal (line a in diagram below). However, under a variation of search strategy there is a tendency for recall to increase as precision decreases, or vice versa, as shown in line b below.



The data indicates that this effect may have occurred so the possibility of a change of strategy of the person making relevance judgments must be considered. When making relevance judgments using only the title of an article, a person may infer that certain questions are relevant to the article represented on the basis of a single word. He may then include many irrelevant questions he guesses to be relevant; as a result, his recall may be high and his precision low. As the person has increasing amounts of information as a basis of his judgments, he may be more cautious which may increase his precision. But, at the same time, he may miss more relevant material than he did when he guessed--this may decrease his recall.

The average recall and precision for each person over all articles judged is shown in figure 1. There does not appear to be a great difference in the results for each person.

The greatest dispersion of points of recall and precision occurred in Part 1. It had, however, a higher average of recall and precision for both the 60 and 300 word abstracts than did Part 2. The average for both the 30 and 60 word abstracts in Part 2 does show a recall and precision trade-off but not to a great extent. All the recall and precision points for each length representation for each person and article are shown in figure 2 (page 14).

It is also of interest to see the percentage of change in recall and precision and the changing nature of the trade-off between them as the length of abstract increases. To this end all lines are moved to a common origin. This is shown in figure 3.

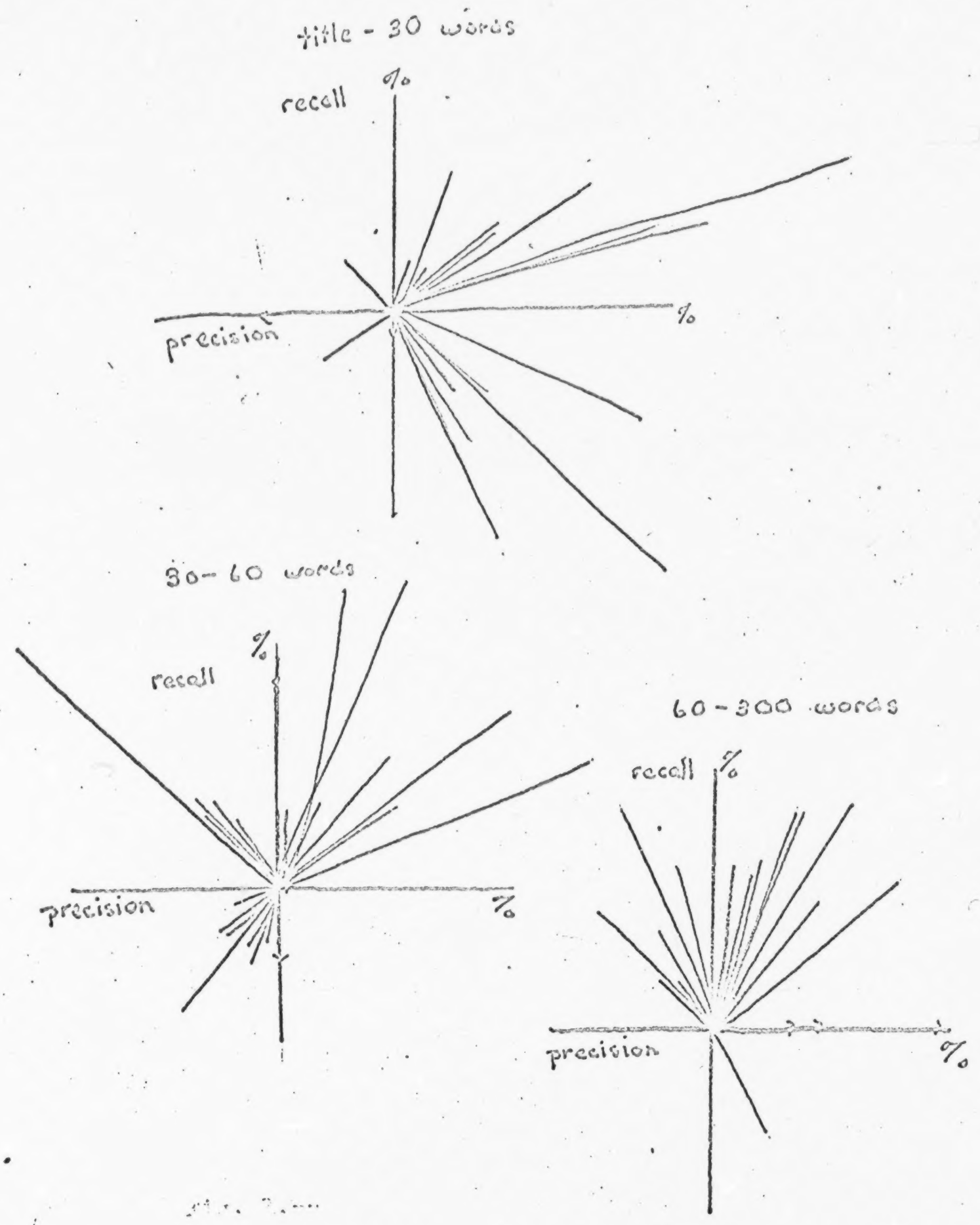


Fig. 3.--Improvement in Recall and Precision for Each Person and Each Length Abstract.

Since all lines are moved to a common origin, the initial point of recall and precision for each case is not indicated. The difference in the amount of change for each case can be compared but not evaluated on an absolute basis. For Parts 1 and 2 combined, there is a significant amount of increase in precision in the majority of cases when one moves from the title to the 30 word abstract, but in almost half the cases recall decreases. From the 30 to the 60 word abstract, there are more cases of significant increase in recall but more than half the cases show a decrease in precision. From the 60 to the 300 word abstract there are again increases in recall, a third of the cases showing a decrease in precision. The increases in recall for 60 to 300 word abstracts are not so great as those which occur in the change from the 30 to the 60 word abstract. This is to be expected as there is less room for improvement as recall and precision approach closer to the ideal point.

The average improvement in recall and precision for each person over all articles for Part 1 and Part 2 is shown in figure 4. There is some variation but also a steady trend in an increase in precision and recall is indicated in most cases. The average improvement over all persons and all articles shows a greater trend to improvement in recall than in precision. The initial points of recall and precision for each person and the average initial point are indicated.

For each move to a longer length abstract, there is at least one case where both recall and precision decrease, called a negative change. From the 30 to the 60 word abstract there are 6 negative changes; neither of the other lengths of abstract have more than two.

In making repeat relevance judgments to the articles for Part 1, on the average, 16 per cent of the judgments did not match the base list of questions. This difference in the matches was called the inconsistency in judgment. Though there is only a 16 per cent difference in matches for the two judgment sets, there may be another 16 per cent of the judgments which match both times only by chance. A consistency of just 68 per cent can be assumed. For the pilot study an inconsistency of 18 per cent was found for the same persons when they made relevance judgments for articles at two different times. The standard deviation of the average inconsistency for each person from the total average inconsistency was 3.8 per cent in the pilot study and 3.2 per cent in Part 1.

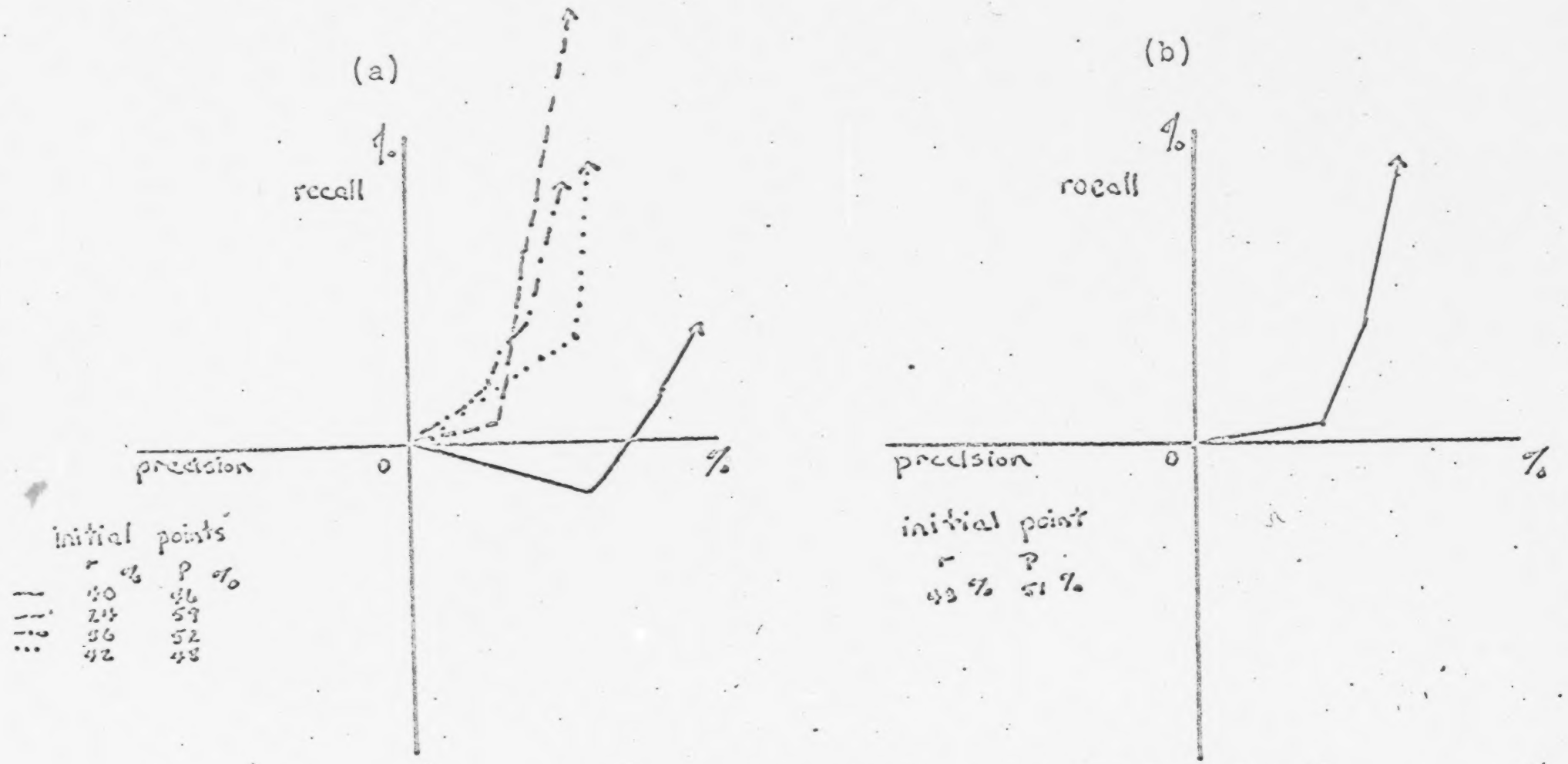


Fig. 4.--Average Improvement in Recall and Precision for Each Length Abstract over All Articles.

(a) For Each Person
(b) For All Persons

The percentage of total correct judgments was quite high and about the same for all judges. The range of the average was from 68 per cent to 76 per cent for title judgments and from 82 per cent to 90 per cent for the 300 word abstract. The standard deviation for title judgments was 3.8 per cent and 1.6 per cent for the 300 word abstracts. As seen, the range of standard deviation is much wider for the title or shortest representation than it is for the longest representation.

Because there was an inconsistency in judging the relevance of questions to an article on two occasions by a particular person, recall was calculated again for each length representation using as the base list only those questions agreed upon as relevant both times by the person. Higher recall was expected; since if the matching questions and their relevance to an article were clear for the judger when deciding relevance for the whole article, most likely they would be less difficult for the person to pick out when judging probable relevance using a representation. The comparison of the average recall over all articles for each person for each method of calculation is shown in Table 1. As can be seen, recall is generally slightly higher using the questions matching both times, but not to a great extent.

DISCUSSION OF RESULTS

There are several possible explanations for recall and precision trade-off or for negative changes. As mentioned, one of these could be a change of strategy in judging a question probably relevant to an article; more gambling or guessing may exist for the title or 30 word abstract, which may by chance be correct. When the person has more information, he may change his mind and perhaps incorrectly find it an irrelevant question. On the other hand, if there were much guessing, a larger number of questions might be expected to be judged relevant for the title than for the 300 word abstracts. But, the average number of questions related using the title was only 21, and for the 300 word abstract, 22. The standard deviation for the title was 7.4 questions and 7.5 questions for the 300 word abstract.

Another possibility is that trade-off and negative changes reflect the cases that are doubtful or difficult to judge. The longer abstract could also contain misleading information on the emphasis or coverage of the article.

Since there is less scattering of recall and precision points for Part 2 than for Part 1, there could be less inconsistency. This could be influenced by differences in the question lists, abstracts, or the articles themselves. Since the abstractors had been exposed to the question list of Part 1 when they judged relevance before they wrote abstracts for the articles in Part 2, this may account for more consistency in Part 2. They may have come to a common understanding of the procedure and what would be expected of them; the question list was a piece of information held in common. It may also be possible that persons become more consistent with each other with repetition of a certain type of task.

In comparison of the results of this study and those of the Rath, Resnick, and Savage study, mentioned on page 3, the title performs quite well both times. With a different measure for evaluation, perfect performance would be 23 per cent of the documents accepted as relevant and 77 per cent of the documents rejected as irrelevant.[1] Use of the title brought 21 per cent correct acceptances (total of 38 per cent acceptances) and 57 per cent correct rejections (total of 62 per cent rejections). With the whole text there was 19 per cent correct acceptances (total of 30 per cent acceptances) and 67 per cent correct rejections (total of 70 per cent rejections). There was some improvement with the use of the two different abstracts, automatically prepared, over the title but the results were quite similar to those for the full text.

---

[1] Rath, Resnick and Savage, American Documentation, XII (April, 1961), 128.

CONCLUSIONS AND SUGGESTIONS

Further experimentation with relevance judgments and how their effectiveness is related to the length of the representations used to make them would be of interest. The study being reported showed that the title alone was quite effective in the making of relevance judgments and that even with a 300 word abstract, this effectiveness, in terms of recall and precision, was not doubled. However, the use of an abstract did increase the recall and precision, but the results of the study do not clearly suggest an optimal length abstract. There could be implications for information systems in research along this line. If the effectiveness of a system can be noticeably improved for the user when he is judging the relevance of a list of articles retrieved in response to his request by the inclusion of some length representation for each article in the list, this would be useful in planning an information system.

There are implications for indexing in attempting to decide which was the "best" length representation of an article. Depending on the recall and precision demanded of a system, a title or an abstract could be used for indexing, rather than the whole article. Indexing of a representatation of an article would be expected to be easier and quicker.

If this experiment or one like it were to be repeated, it would be interesting to see the effect of some variables not considered in the study reported.[1] Some of the inconsistency could be related to the yes or no type of relevance judgment for Parts 1 and 2 that the judgers were asked to make. The influence of more available information about an article and of shifts in strategy as abstract length increases might be reflected if persons were asked to rate their relevance judgment on a five or more point scale of relevance. The number of articles, abstractors and judgers could be increased and provide a larger sample size and thus more reliable data. Other lengths of abstracts should be tested to see if judgments of relevance are more effective, for example, a 100 word abstract may perform better than the 60 word abstract although not as well as the 300 word abstract.

Perhaps, different persons should perform the tasks of abstracting and judging relevance. Although the weakness of the pilot study in the number of questions and their direct relation to the articles was corrected in Parts 1 and 2, the judgers may have felt some unconscious hesitation in relating questions which

---

[1] A list of variables influencing relevance judgments is contained in the System Development Corporation, Experimental Studies of Relevance Judgments, Second Progress Report prepared for National Science Foundation, TM-3068/000/00, July 1, 1966.

they knew were answered by their article to the other articles. As a control for the question lists, one of the lists should contain no questions formulated from the articles to be judged.

Two other types of representation of an article might be included. Since caption and table headings are sometimes quite informative about the article, they might be judged along with the title after completion of title judgments. A list of index terms assigned to the article, approximately the same length as the number of substantive words in a 30 word abstract, could be judged before the 30 word abstract. This might give some indication of advantage or disadvantage of a connected discourse.

When analyzing the data, two methods for determining recall and precision could be used. The calculation of recall and precision could be done twice for each length representation of an article. The base list of questions and the repeat judgments could be used as the two standards to measure the number of questions correctly judged relevant using a representation of an article. Then, the recall and precision should be compared to see if there is any difference.

Or, only those questions whose judgments matched on the two lists of judgments made using the entire article as a basis could be compared to the questions judged probably relevant on the basis of the representations. This will result, most likely, in a shorter list of questions.

In conclusion, as expected, recall and precision did increase as the length of abstract increased. The amount of the increase was not, however, as much as was expected. Trade-off of recall and precision, i.e., as recall increased, precision decreased or vice-versa, did occur when the length of abstract changed but not in every case. Increasingly longer abstracts did provide useful information to judge relevance; although the 300 word abstract did not double the effectiveness of the title, there was a significant increase in recall and precision, of about 50 per cent in both, over the recall and precision for the title. Moveover, the number of total correct judgments consistently increased as the length of abstract increased.

The expectation that recall and precision would increase to a great extent with the addition of five times as many words to the title was not realized. Increase in recall and precision does not seem to be a function of length of representation. As the basis for relevance judgment, the use of the title alone could be justified. However, use of an abstract with the title was shown to result in improvement significant enough to warrant consideration.

APPENDIX

TABLE 1

AVERAGE RECALL FOR EACH PERSON OVER ALL ARTICLES

	Length of Representation	Part 1 Per Cent	Recalculated Using Only Those Questions Matching Per Cent
	Title	42	45
Judge B	30	38	43
	60	43	46
	300	60	65
	Title	23	30
*Judge D	30	40	48
	60	34	44
	300	63	74
	Title	52	52
Judge E	30	47	55
	60	51	52
	300	72	68
	Title	55	55
Judge K	30	63	69
	60	63	71
	300	79	81

*One article.

TABLE 2

PART 1: RECALL---PRECISION

Judge B	Length of Representation	Article 1 Recall Per Cent	Article 1 Precision Per Cent	Article 2 Recall Per Cent	Article 2 Precision Per Cent	Article 4 Recall Per Cent	Article 4 Precision Per Cent
	title	8	23	47	41	55	46
	30	15	25	58	79	40	80
	60	31	67	68	87	30	75
	300	62	54	68	87	50	100
Judge D		Article 1		Article 2		Article 3	
	title	25	50	23	67	23	61
	30	21	45	12	75	40	88
	60	62	62	62	84	34	80
	300	83	71	69	76	63	92
Judge E		Article 2		Article 3		Article 4	
	title	37	67	68	78	50	58
	30	44	60	50	88	47	58
	60	56	50	61	89	37	58
	300	74	65	75	81	67	77
Judge K		Article 1		Article 3		Article 4	
	title	50	30	57	63	59	77
	30	44	21	67	77	79	85
	60	67	52	60	45	62	72
	300	83	36	60	86	93	84

TABLE 3

PART 2: RECALL--PRECISION

		Article 1		Article 2		Article 3	
Judge B	Length of Representation	Recall Per Cent	Precision Per Cent	Recall Per Cent	Precision Per Cent	Recall Per Cent	Precision Per Cent
	title	22	33	35	83	70	52
	30	33	75	35	64	35	88
	60	44	80	46	80	70	48
	300	50	75	54	82	70	78
Judge E		Article 2		Article 3		Article 4	
	title	38	45	68	30	76	65
	30	42	62	37	44	65	79
	60	50	54	47	33	62	72
	300	33	53	32	40	62	86
Judge K		Article 1		Article 3		Article 4	
	title	29	31	44	47	15	38
	30	35	35	56	62	35	100
	60	53	50	50	53	44	94
	300	76	56	72	48	68	92

TABLE 4

PART 1: TOTAL CORRECT JUDGMENTS

		Length of Representation				
B	Article	Title	30	60	300	Per Cent Increase from Title--300
	100	65	83	89	89	24
	100	77	88	91	91	14
	100	77	86	85	90	13
	average	73	86	88	90	
D	100	69	73	81	87	18
	100	74	69	87	85	11
	100	62	75	60	82	20
	average	68	72	76	84	
E	100	98	77	74	82	4
	100	86	84	87	88	6
	100	74	73	73	83	9
	average	78	78	78	84	
K	100	78	71	81	81	2
	100	74	81	79	81	7
	100	76	79	83	84	8
	average	76	77	81	82	

TABLE 5

INCONSISTENCY IN JUDGING RELEVANCE OF QUESTIONS
USING WHOLE ARTICLE AS BASIS

Pilot Study	Per Cent	Per Cent	Per Cent	Average Per Cent
B	6	27	22	18
D	13	38	25	25
E	12.5	21	26	20
K	15	15	17	16

Total average--18 per cent

Part 1:

B	8	8	18	11
D	14			14
E	14	16	24	18
K	19	18	19	19

Total average--16 per cent

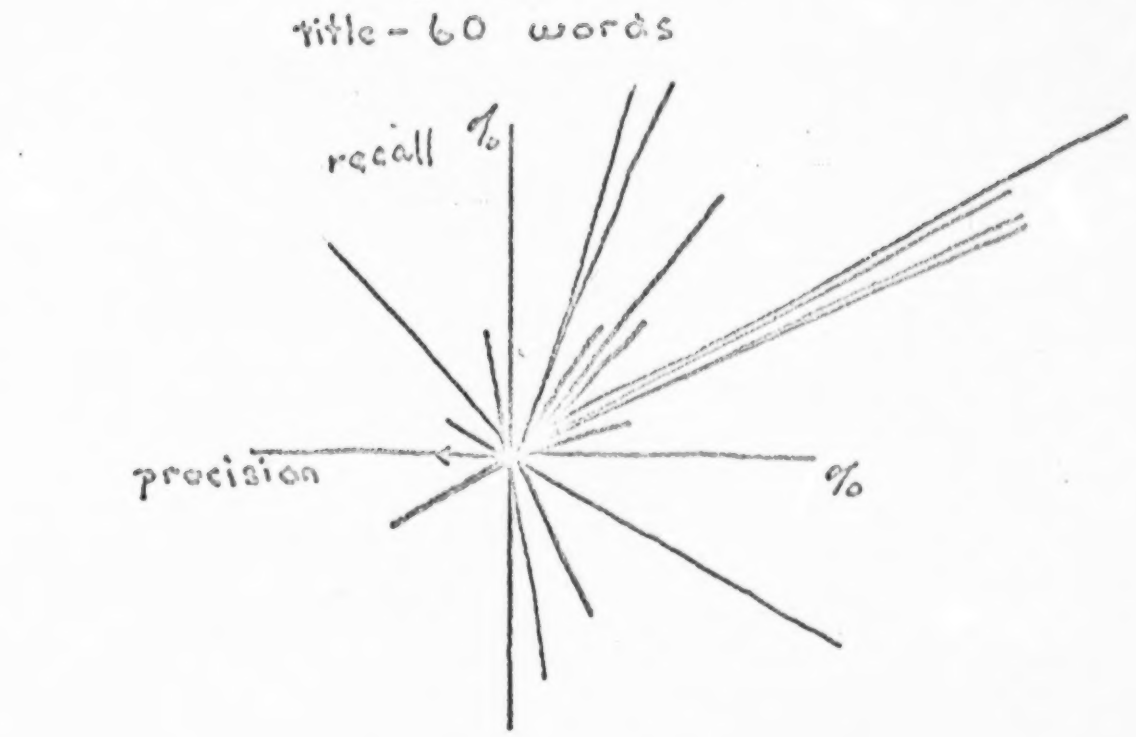


Fig. 5.--Cumulative Improvement in Recall and Precision for Each Person for Title--60 Words and Title-300 Words.

ARTICLES SELECTED FOR USE IN STUDY

Pilot Study

1. Chapin, R. E., and Shilling, C.W. "A Model for the Study of Scientific Communication," American Documentation, XIII (October, 1962), 410-13.

2. Garvey, W. D., and Griffith, B. C. "Scientific Information Exchange in Psychology," Science, CXLVI (December, 1964), 1655-59

3. Schultz, Louise. "The Information System; Too Big and Growing," American Documentation, XIII (July, 1962), 288-94.

4. Vickery, B. C. "Scientific Information: Problems and Prospects," Minerva, I (January, 1963), 21-38.

Part 1

1. Bar-Hillel, Yehoshua. "Is Information Retrieval Approaching a Crisis?" American Documentation, XIV (April, 1963), 95-98.

2. Feinler, E. J., et al. "Attitudes of Scientists toward a Specialized Information Center, "American Documentation, XVI (October, 1965), 329-33.

3. Swanson, Don R. "On Improving Communication among Scientists," Bulletin of Atomic Scientists, XXII (February, 1966), 8-12.

4. Weinberg, Alvin. "Second Thoughts on Scientific Information," College and Research Libraries, XXV (November, 1964), 463-71.

Part 2

1. Leake, Chauncey D. "What Must Give in the Documentation Crisis?" Information Retrieval Management, ed. Lowell H. Hattery and Edward M. McCormick (Detroit, Michigan: American Data Processing, 1962), pp. 15-20.

2. Shilling, Charles. "Requirements for a Scientific Mission-Oriented Information Center," American Documentation, XIV (January, 1963), 49-53.

3. Swanson, Don R. "Information Retrieval: State of the Art," Proceedings of Western Joint Computer Conference, May 9-11, 1961 (Los Angeles, Calif.: Western Joint Computer Conference, 1961), XIX, 239-46.

4. Weinberg, Alvin. "Scientific Communication," International Science and Technology, II (April, 1963), 65-74.

Article, Not Abstracted

Dowe, Ray M., Jr. "Collection and Dissemination of Scientific Information," Military Review, XLIV (January, 1964) 55-65.

BIBLIOGRAPHY

DeLucia, Al. "Index-Abstract Evaluation and Design," American Documentation, XV (April, 1965), 121-25.

Doyle, Lauren. "Is Relevance and Adequate Criterion in Retrieval System Evaluation?" Automation and Scientific Communication, Part 2 (1963), pp. 199-200.

Hirayama, Kenzp. "Length of Abstract and Amount of Information," Journal of Chemical Documentation, IV, No. 1 (January, 1964), 9-11.

Montague, Barbara A. "Testing, Comparison and Evaluation of Recall, Relevance and Cost of Coordinate Indexing with Links and Roles," American Documentation, XVI (July, 1966), 201-208.

Rath, G. J., Resnick, A., and Savage, T. R. "Comparison of Four Types of Lexical Indicators of Content," American Documentation, XII (April, 1961), 126-30.

Rees, Alan M., and Saracevic, Tefko. The Measurability of Relevance. A paper published by the Center for Documentation and Communication Research of Western Reserve University, October, 1966.

Resnick, A. "Relative Effectiveness of Document Titles and Abstracts for Determining Relevance of Documents," Science, CXXXIV (October 6, 1961), 1004-1006.

Resnick, A., and Savage, T. R. "The Consistency of Human Judgments of Relevance," American Documentation, XV (April, 1965), 93-95.

Salton, Gerald. "The Evaluation of Automatic Retrieval Procedures-Selected Test Results Using the SMART System," American Documentation, XVI (July, 1966), 208-222.

Swanson, Don R. "Searching Natural Language Text by Computer," Science, CXXXII (October 21, 1960), 1099-1104.

System Development Corporation. Experimental Studies of Relevance Judgments. Second Progress Report. A report prepared for the National Science Foundation, TM 3068/000/00, July 1, 1966.

Taube, Mortimer. "A Note on the Pseudo-Mathematics of Relevance," American Documentation, XVI (April, 1966), 69-73.

Taylor, Robert. "The Process of Asking Questions," American Documentation, XIII (October, 1963), 391-96.

Articles Selected for Use in Study

Pilot Study

1. Chapin, R. E. and Shilling, C. W. "A Model for the Study of Scientific Communication," American Documentation, XIII (October, 1962), 410-13.

2. Garvey, W. D., and Griffith, B. C. "Scientific Information Exchange in Psychology," Science, CXLVI (December, 1964), 1655-59.

3. Schultz, Louise. "The Information System: Too Big and Growing," American Documentation, XIII (July, 1962), 288-94.

4. Vickery, B. C. "Scientific Information: Problems and Prospects," Minerva, I (January, 1963), 21-38.

Part 1

1. Bar-Hillel, Yehoshua. "Is Information Retrieval Approaching a Crisis?" American Documentation, XIV (April, 1963), 95-98.

2. Feinler, E. J., et al. "Attitudes of Scientists toward a Specialized Information Center," American Documentation, XVI (October, 1965), 329-33.

3. Swanson, Don R. "On Improving Communication among Scientists," Bulletin of Atomic Scientists, XXII (February, 1966), 8-12.

4. Weinberg, Alvin. "Second Thoughts on Scientific Information," College and Research Libraries, XXV (November, 1964), 463-71.

Part 2

1. Leake, Chauncey D. "What Must Give in the Documentation Crisis?" Information Retrieval Management, ed. Lowell H. Hattery and Edward M. McCormick. Detroit, Michigan: American Data Processing, 1962. Pp. 15-20.

2. Shilling, Charles. "Requirements for a Scientific Mission-Oriented Information Center," American Documentation, XIV (January, 1963), 49-53.

3. Swanson, Don R. "Information Retrieval; State of the Art," Proceedings of Western Joint Computer Conference, May 9-11, 1961. Los Angeles, Calif.: Western Joint Computer Conference, 1961. XIX, 239-46.

4. Weinberg, Alvin. "Scientific Communication," International Science and Technology, II (April, 1963), 65-74.

Article, Not Abstracted

Dowe, Ray M., Jr. "Collection and Dissemination of Scientific Information," Military Review, XLIV (January, 1964), 55-65.